CSC110 Final Project : The Effects of COVID-19 on Community Mobility in Canada

Adrian Hui, Brandon Tang, Jillian Soriano, Nzube Ogbu

December 14, 2021

Problem Description and Research Question

We are all too familiar with the countless lock downs and restrictions following the beginning of the COVID-19 pandemic. All around the world, governments have enacted policies to restrict public mobility.
In Canada for example, in order to board VIA Rail, Rocky Mountaineer trains, and domestic or international flights for most airports, it is required for those twelve years of age (plus four months) or more to be fully vaccinated. (Canada, 2021) In addition, Canadian travellers need to have their second dose at least fourteen days before travel, have no COVID-19 symptoms, and must wear a mask. Although it is still possible to travel with these regulations, these regulations undeniably limit the amount of travel occurring.

Evidently, the implementation of such rules differs from country to country and even city to city. Given that U of T is a diverse school with students from a variety of countries, apparent even within our own group, we were curious how our different experiences during the pandemic translated in terms of mobility and if this could offer insight on differences in government policies and other lurking variables. After all, our freedom to move orchestrates our daily lives. Hence, our research question is: **How did the pandemic impact local travel within Canada and did the change in mobility vary depending on the province or territory?** To answer this question, we will do a continuous comparison of mobility trends over the course of the pandemic.

Analyzing community mobility is essential during this pandemic because it allows government officials to design effective policies that minimize risk of COVID-19. We can quantify change in mobility by measuring the time spent at or the number of visitors at a location. In essence, the traffic or accessibility of various locations reflects mobility. Certainly,mobility changed with varying degrees depending on region, from whole province shut downs to local scale trips to the grocery stores. Change in mobility may reveal insight into the well-being of communities and its citizens as well as reflect government policy, wealth, and other variables. For instance, poorer neighbourhoods where most jobs consist of physical labor may see minimal change in workplace mobility while wealthier communities where most people work in corporate offices and have plenty of access to technology at home may see a major decrease in workplace mobility. Consequently, COVID-19's effects on mobility are intertwined with other factors and those other factors must be considered when analyzing mobility during the pandemic. This project aims to explore the possible magnitude of these restrictions on mobility, the diversity of its effects and how they differ from province to province.

Dataset Description

The dataset we are going to be using, is the Community Mobility Reports created by Google using data collected from Google Maps, government data, and other Google services. The reports chart movement trends over time by geography, focusing on six categories: retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. The data is available in a CSV file that includes various countries and their regions alongside the percent change in the number of visitors for categories of place, relative to the median-day value from January 3, 2020 to February 6, 2020. It is important to note that there exists some gaps in data when the data does not ensure anonymity and privacy of community members. According to Google, these gaps do not necessarily mean that no mobility occurred on that day, they should just be treated as true unknowns. It is important to note that for Prince Edward Island, the dataset's data only includes data for Prince Edward County, not the province as a whole. In our report, We will be focusing on Canada, subdividing the data into the individual provinces and territories.

Our processed dataset version is also in the csv format. We cleaned the original 2020_CA_Region_Mobility_Report" from the raw dataset, making a few changes to better base our program on; the edited dataset does not have any sub regions within the provinces or any rows that are missing one or more pieces of data (except for the territories and prince edward). We used columns: B (for the province name), I (for the date), J, K, L, M, N, and O (for the different percentages for each type of mobility category). Since this data is not solely dependent on the pandemic, but also other influences (e.g. national and local holidays, wealth, accessibility, types of labour, etc), we will further investigate each province and territory and seek other influential factors. This will give insight on how external circumstances cause each area to differ.

0.1 Instructions for obtaining Datasets

This is the link of the raw dataset of Canada: `https://www.google.com/covid19/mobility/`. To obtain the raw dataset with Canada's data, access this link, download the Region CSVs. In the Region CSVs folder, Canada's dataset is named "2020_CA_Region_Mobility_Report".

In order to access the processed version of the data, visit `https://send.utoronto.ca/` and pick up the file with the following information:
- Claim ID: afZSFUv7aR4ddh3D
- Claim Passcode: eAJhArF3pJTmcJVW

Computational Overview

0.2 Describe the major the computations your program performs, such as: data transformation/filtering/aggregation, computational models, and/or algorithms.

We first transformed the data into a pandas dataframe using the pandas library. Then we manually filtered the data to include the exact details we wanted, namely the percent change from baseline of mobility for each province throughout the course of the 2020-2021 year. We then filtered the data by column to create our Province object class which stored all the mobility data from the dataset for that specific province as well as did several computations. The Province class has methods average() that finds the average of a mobility category and plot_data() which generates a scatter plot of a province's or a territory's percentages in a mobility category. We also created several functions for visualizing data. The lin_reg() creates a scatter plot using the mobility data of the user's choice and performs a linear regression while plotting a line of best fit onto the scatter plot. The regression uses a train-test split of 0.33 and we used mean squared error, the average squared distance of each point from the line of best fit, to judge how well the line fit the data. Additionally, k_means() performs k-means clustering on two mobility categories of two different provinces with 1 to 8 centroids, and plots the best result of the clustering. The ideal amount of centroids was determined by plotting an elbow chart with the inertia of the clusterings and prompting the user to choose their desired amount of clusters which was done in the find_elbow function. Again, inertia is essentially the sum of squared distances of each point to its centroid and it provides a way to measure the effectiveness of the clustering. Additionally, bar_graph() and bar_graph_two() also compute the means of the derided change in mobility percentage for every province and plots that data. Finally, we calculated the z-score of a mobility score compared to the average of percentages of the category it belongs to. using the calculate_z_score() function. To calculate the z-score, the mean of the percentages in a mobility category from a data point is calculated. A mobility percentage in the same mobility category is then subtracted by the mean. This difference is divided by the standard deviation of the percentages in the mobility category.

0.3 Explain how your program reports the results of your computation in a visual and/or interactive way.

When main.py is run, it gives the user multiple prompts. Where each answer per prompt is used as an argument for a function. The first prompt asks for a province in which the answer is stored in the variable user_plot_choice and will be passed as an argument or a key to access certain data to plot_data(), k-means(), and histogram(), and data_picker() for variables plot_x and plot_y. It is also used for bar_graph_two(). The second prompt also asks for a province in which the answer is stored in the variable user_plot_choice2 and is used as a key in an argument for k-means(). The third prompt asks for a mobility category that will be stored in user_data_choice1 and will be passed as an argument or a key to access certain data to plot_data(), lin_reg(), k_means(), histogram(), and data_picker() for variable plot_x. The fourth prompt asks for another mobility category that is stored in user_data_choice2 and is used as an argument for lin_reg() and data_picker for variable plot_y. User_data_choice2 is also used as a key for an argument in k_means). The fifth prompt asks for a mobility category that is stored in hist_choice and is used as an argument for histogram() and it is the mobility category that the calculate_z_score function() will base its results on. The sixth prompt asks for how many k clusters the user wants for the k_means clustering model. The last prompt asks the user for a percentage that they would like to compare to the average percentages of the mobility category specified in the fifth prompt.

After all of the prompts are answered, the program generates a scatter plot, a linear regression model, a k-means distortion model, a k-means clustering model, a histogram, and two bar graphs. The program also generates a z score that compares a mobility percentage to a province's percent average for a specified mobility category. By taking user input, the visuals and information displayed is customized to the interests of the user, allowing them to analyze mobility for which every province and category they desire.

0.4 Explain how your program uses new libraries to accomplish its tasks. Refer to specific functions, data types, and/or capabilities of the library that make it relevant for accomplishing these tasks.

Our program uses four new libraries:

1. Numpy: We use numpy to create numpy arrays which make it more convenient to plot and compute on data for the k-means clustering. We are essentially able to combine separate columns of data into (x, y) value pairs.

2. Pandas: We mainly use Pandas to access and transform pieces of data in our CSV file. for instance, various functions such as read csv() and Dataframes.index() will allow for quick and easy manipulation of data, creating our own table directly from the csv file. Also,In the calculate_z_score function, Pandas' standard deviation function and mean function is used to calculate the z score.

3. Scikit_learn: We used scikit_learn to implement K-means clustering. K-means is a clustering algorithm that essentially chooses n centroids or centers and groups surrounding data points based on mean distance. This should reveal certain patterns or trends that are overlooked in the data. Ideally, the counties or regions will be clustered by their province. This will also provide us a visual interpretation of how the data is grouped.

4. Matplotlib: Matplotlib is a fundamental library we use to visualize our data. It is used for our bar graphs, linear regression model, and histogram (bar_graph(), bar_graph_two(), lin_reg(), k-means(), histogram()).

Instructions for running the program

In order to run our program, first download all of the MarkUs files and save them under the same directory. Then download the processed version of the dataset and save it in the same directory as the MarkUs files. Then run the entirety of main.py and answer each automated prompt with provinces/territories and mobility categories, etc of your own choosing. After the prompts, models and statistics should be generated relative to the answers given.

Descriptions of changes to the project plan between the proposal and the final submission

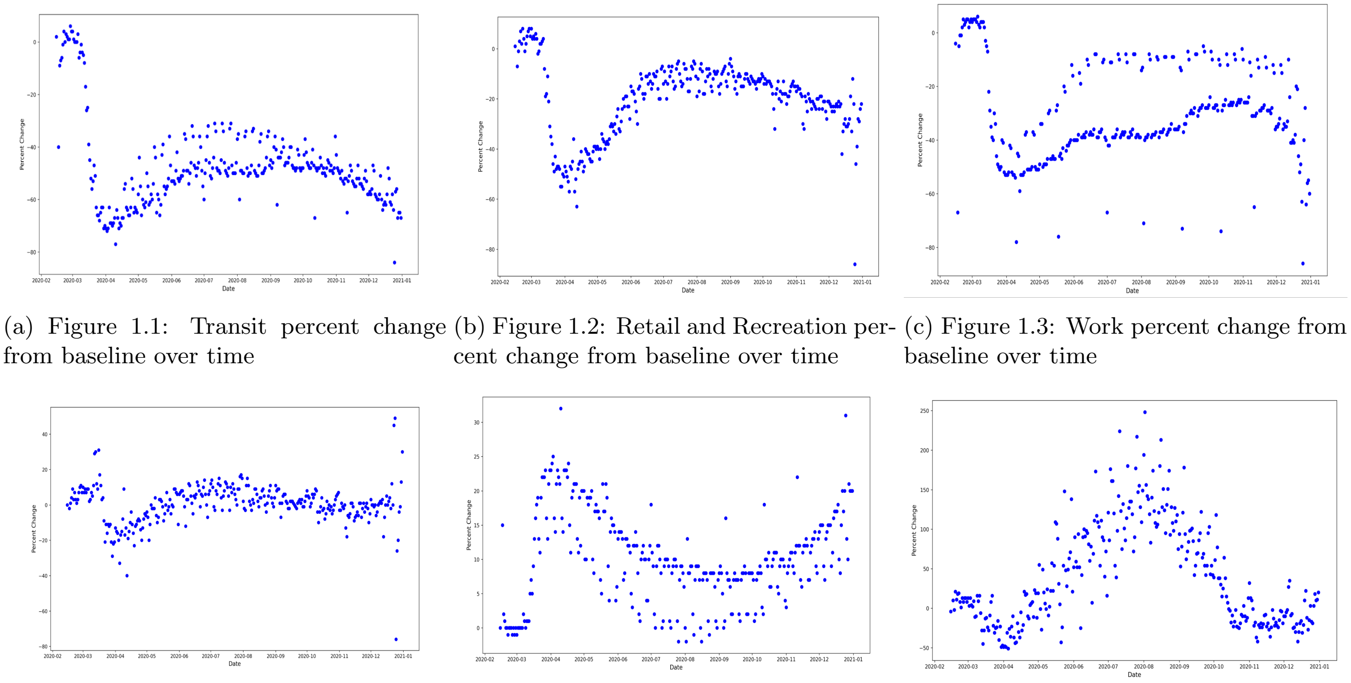
Originally, we planned to analyze the mobility percentages in each province and then the percentages in the sub regions within each province. However, we realized that it would be too impractical and unrealistic to work on that much data. Thus, we limited the data we used to just the percentages of each province as a whole. In general, we wanted to minimize the amount of unwanted data as possible in our dataset. We removed rows containing subregions and Nan values and we removed columns A, C, D, E, F, G, and H because of their irrelevance to our program. Ultimately, this left us with the columns containing the province names, the dates throughout the 2020 - 2021 year, and the percentages of each mobility category; where each percentage corresponds to a date.

Further, In our proposal, we wanted to use pandas, scikit-learn or Tensorflow, and plotly in our program. In the end, we did not use Tensorflow and plotly. We used scikit-learn over Tensorflow because our proposal feedback said that Tensorflow was more complex to use. As for plotly, we were much more comfortable using matplotlib.

Discussion and Conclusion

0.5 Do the results of your computational exploration help answer this question?

Recall that our question was: **how did the pandemic impact local travel within Canada and did the change in mobility vary depending on the province or territory?** Also recall that the percentages in the CSV file are how much the day value has changed compared to the median-day value from January 3, 2020 to February 6, 2020. In order to answer our question, we sought out certain patterns in the data that might reveal trends about mobility throughout Canada. We used the functions plot_data() and data_picker() to generate a scatter plot for a province or territory; which maps a date to its percent change in a mobility category. In our findings, we have noticed that the mobility fluctuated significantly over time. For example, this is seen with Alberta and the mobility categories in *Figures 1.1 to 1.6.* Evidently, these patterns were a result of various lockdowns and COVID-19 protocols enacted throughout the 2020 year. More specifically, we saw that generally across provinces for the Transit, Grocery, Work, Retail and Recreation categories, the percentages dipped during the months of March, April, and May and started rising at the beginning of June before dropping near the end of 2020. The drop coincides with the beginning of the pandemic and the amount of lockdown restrictions being implemented. The novelty of the virus and public fear of it could also be an explanation for this drop. People did not want to go outside as the virus was very new and there was little information about the effects of it. Although there was a slight drop in percentages during the same months for the Parks category, the percentages were generally very high (compared to the other categories) from June to near the end of 2020. In the residential category, in contrast to the other categories, there was a rise in percentages during March, April and May. The category saw a decrease of percentages at the beginning of June before rising again near the end of 2020. The rise in percentages is the result of people staying home because of lockdown restrictions. The drop could be because of the decrease in public fear as more information was being released on Covid-19 and the need for some people to go to work. While the second rise in percentages could be due to the holiday seasons. Even though there was a general pattern across provinces, the amount of change in percentages varied from province to province.



(a) Figure 1.1: Transit percent change from baseline over time

(b) Figure 1.2: Retail and Recreation percent change from baseline over time

(c) Figure 1.3: Work percent change from baseline over time

(d) Figure 1.4: Grocery percent change from baseline over time

(e) Figure 1.5: Residential percent change from baseline over time

(f) Figure 1.6: Parks percent change from baseline over time

Figure 1: Alberta's Percentages for Each Mobility Category Over Time

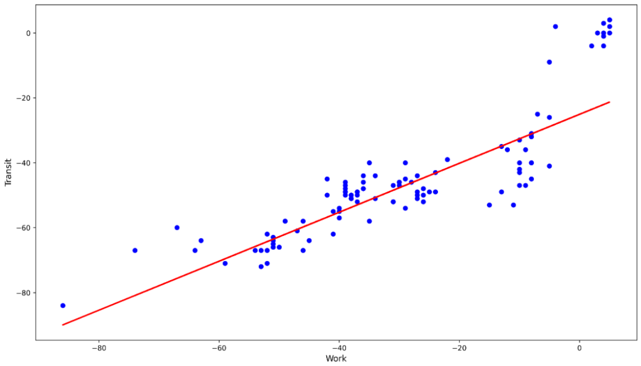


Figure 2: Transit versus Workplace mobility percent change from baseline in BC

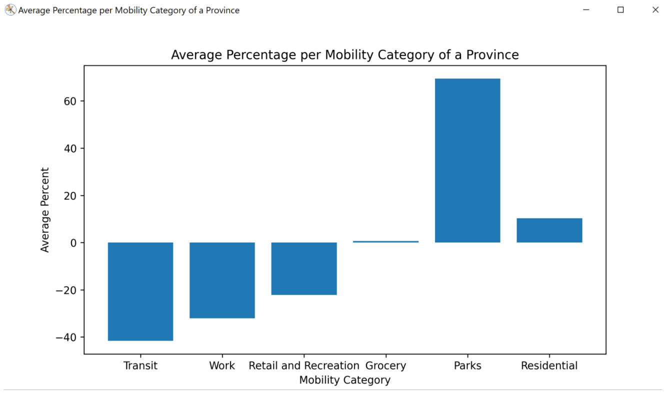

Average Percentage per Mobility Category of a Province
Average Percentage per Mobility Category of a Province
60
40
20
0
−20
−40
Average Percent
Transit
Work
Retail and Recreation
Grocery
Parks
Residential
Mobility Category

Figure 3: Average mobility percent change from baseline in BC

We also performed a linear regression between the various categories of data. Unsurprisingly, we found that certain mobility categories were moderately to highly correlated with a lower MSE. For instance, as seen in *Figure 2*, transit and work mobility seem to be correlated. This makes sense as the majority of public transit users are likely people getting to work. In most cases, the correlation appeared rather quadratic or exponential rather than linear. This seemed to indicate that the effect of one category, for instance, transit was impacted disproportionately by the mobility change in workspaces in British Columbia. Perhaps, without enough people using public transit or driving, various transportation services shut down as maintenance costs were too high or because of government protocols. We found this pattern to be generally true for all provinces, revealing the impacts of COVID were generally no different from province to province. Evidently, under one country, the reaction or behaviour of the government from province to province should not vary too significantly and the change in mobility reflects this. The results of k-means clustering on different categories of mobility were relatively ambiguous. There was no obvious or intriguing pattern found by observing the clustering. It can be inferred that poor mobility in one category was grouped with poor mobility in the other category as observed in *Figure 3*. This could have been because certain categories were highly correlated with others. We also thought that it was possible that the clusters were really just centered around the dates or time of the data recorded and that might highlight the circumstance revolving around government restriction and social norms like fear of the virus or vaccination at the time. When plotting the average percent change in mobility from the baseline for all provinces and territories throughout Canada, we noticed interesting patterns. For instance, transit mobility decreased most significantly on average in more densely polluted provinces such as Ontario and Quebec. However, when park mobility increased most in less densely populated provinces like Nova Scotia or Newfoundland. It is possible that in provinces like Nova Scotia, there are more parks since it is less densely populated which allowed for more people to enjoy such spaces when other areas were restricted while in Ontario, parks are more rare in major cities like Toronto and many public spaces including parks were shut down. Likewise, Toronto likely accounts for a large portion of the transit mobility score and any change in its mobility would greatly impact the overall mobility of Ontario. Small towns in Nova Scotia likely did not have this same effect. Even more so, it is likely that people living in major cities like Toronto are wealthier, working in offices and other jobs that are easy to transform to remote while jobs in Nova Scotia require individuals to work in-person to provide their goods and services. For instance, Nova Scotia is one of Canada's largest seafood exporters and fishing certailyn cannot be done remotely. Thus, use of transportation to go fishing was not impacted much by the pandemic. We also saw that most mobility categories were normally distributed throughout Canada as seen in *Figure 4* which is not a surprise given that most natural phenomena happens to be normally distributed. Finally, when examining the average percent change for each mobility score for a given province, transit and work mobility were typically most affected during the pandemic. On the other hand, grocery mobility stayed relatively the same as seen in *Figure 5* which makes sense there are no other alternatives to eating whereas work shifted to online and remote. The results of the z-score can easily vary and it is difficult to make any conclusions about a province or territory as a whole based on the z-score of a data point. Overall, it is clear the pandemic impacted mobility in each province slightly differently based on various factors including population density. However, altogether, each province showed similar and cohesive changes in mobility.

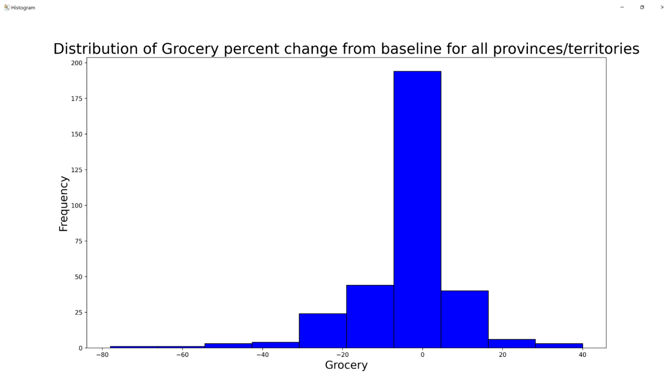


Figure 4: Distribution of grocery mobility percent change from baseline across Canada

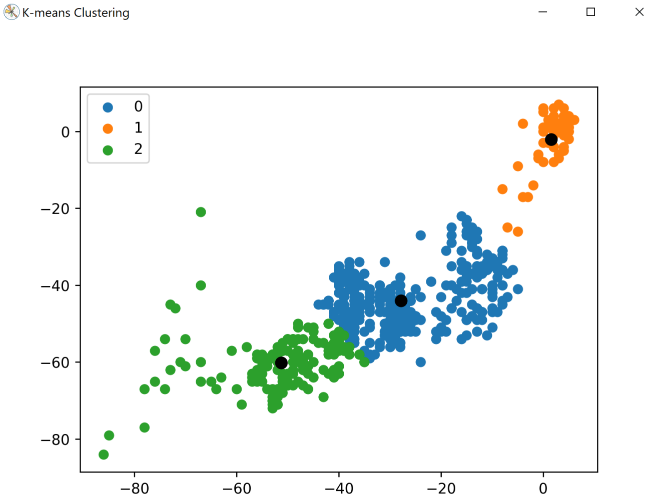


Figure 5: BC Transit versus AB Workplace mobility percent change from baseline

0.6 What limitations did you encounter, with the datasets you found, the algorithms/libraries you used, or other obstacles?

The main obstacles we faced revolved around the original dataset which was very large and difficult to work with. At first, we used the original dataset while writing the program. Since there was data of each of the provinces' and territories' sub regions and counties, it was difficult to create our program so that it could account for all the data. We soon processed the data so that the dataset excluded the data involving the sub regions. Some of the data was also incomplete or missing and without a lot of data points, so the regression model was not as accurate as it could have been. Especially since we were using a train-test split of 0.33, it was clear from the scatter plot that there was not a lot of data available to test our regression on. We were able to remove rows with NaN values for most of the provinces and territories, but we were forced to keep some. This is because some rows contained NaN values and actual data and we did not want to lose a significant amount of data. Also, we had to take some time to read and learn the documentation of several libraries that we were unfamiliar with including the scikit.

0.7 What are some next steps for further exploration?

Next steps might include cleaning up the data or looking to expand the amount of usable data. It would be interesting to see if the mobility patterns we found in Canada would apply to other countries as well or if they would be drastically different. Furthermore, it would be interesting to do more research on the wealth and demographic of each province and see how these things could influence movement trends. We could focus on the movement trends of specific groups of people, analyzing the percentages of mobility categories relative to wealthy people or relative to elderly people. For example, if the focus was on elderly people, there may be less drastic changes to the percentage of people going to work.

References

Bar Charts in Python. Plotly. (n.d.). Retrieved November 5, 2021, from https://plotly.com/python/bar-charts/

Canada, G. A. (2021, November 4). *Covid-19 boarding flights and trains in Canada.* – Travel restrictions in Canada – Travel.gc.ca. Retrieved November 5, 2021, from https://travel.gc.ca/travel-covid/travel-restrictions/domestic-travel

DataFrame — pandas 1.0.3 documentation. (2014). Pydata.org.from https://pandas.pydata.org/docs/reference/frame.html

Google. (n.d.). *Covid-19 Community mobility reports.* Google. Retrieved November 5, 2021, from https://www.google.com/covid19/mobility/

Google. (n.d.). *Understand the data - community mobility reports help.* Google. Retrieved November 3, 2021, from https://support.google.com/covid19-mobility/answer/9825414?hl=en&ref_topic=9822927

Histograms in Python. Plotly. (n.d.). Retrieved November 5, 2021, from https://plotly.com/python/histograms/

Jeon, T. (2018, May 9). *Importing data with pandas' read_csv().* DataCamp Community. Retrieved November 5, 2021, from https://www.datacamp.com/community/tutorials/pandas-read-csv

Kovalev, S., Sintsov, S., & Khizhniak, A. (2019, June 8). Implementing K-means clustering with tensorflow. Altoros. Retrieved November 5, 2021, from https://www.altoros.com/blog/using-k-means-clustering-in-tensorflow/

1.4. Support Vector Machines. scikit. (n.d.). Retrieved November 5, 2021, from https://scikit-learn.org/stable/modules/svm.html#regression

2.3. Clustering. scikit. (n.d.). Retrieved November 5, 2021, from https://scikit-learn.org/stable/modules/clustering.html#clustering

Z-Score: Definition, Formula and Calculation. (n.d.). Statistics How To. fromhttps://www.statisticshowto.com/probability-and-statistics/z-score/